

RESEARCH

Benchmarking network propagation methods for disease gene identification

Sergio Picart-Armada^{1,2*}, Steven J. Barrett⁴, David R. Willé⁴, Alexandre Perera-Lluna^{1,2}, Alex Gutteridge³ and Benoit H.⁸ Dessailly⁵

Abstract

Background: In-silico identification of potential disease genes has become an essential aspect of drug target discovery. Recent studies suggest that one powerful way to identify successful targets is through the use of genetic and genomic information. Given a known disease gene, leveraging intermolecular connections via networks and pathways seems a natural way to identify other genes and proteins that are involved in similar biological processes, and that can therefore be analysed as additional targets.

Results: Here, we systematically tested the ability of 12 varied network-based algorithms to identify target genes and cross-validated these using gene-disease data from Open Targets on 22 common diseases. We considered two biological networks, six performance metrics and compared two types of input gene-disease association scores. We also compared several cross-validation schemes and showed that different choices had a remarkable impact on the performance estimates. When seeding biological networks with known drug targets, we found that machine learning and diffusion-based methods are able to find novel targets, showing around 2-4 true hits in the top 20 suggestions. Seeding the networks with genes associated to disease by genetics resulted in poorer performance, below 1 true hit on average. We also observed that the use of a larger network, although noisier, improved overall performance.

Conclusions: We conclude that machine learning and diffusion-based prioritisers are suited for drug discovery in practice and improve over simpler neighbour-voting methods. We also demonstrate the large effect of several factors on prediction performance, especially the validation strategy, input biological network, and definition of seed disease genes.

Keywords: guilt by association; biological networks; network propagation; drug targets; genetic associations; cross validation; protein complexes

*Correspondence:

sergi.picart@upc.edu

¹B2SLab, Departament d'Enginyeria de Sistemes, Automàtica i Informàtica Industrial, Universitat Politècnica de Catalunya, CIBER-BBN, 08028, Barcelona, Spain

Full list of author information is available at the end of the article

†Equal contributor

Background

The pharmaceutical industry faces considerable challenges in the efficiency of commercial drug research and development [1] and in particular in improving its ability to identify future successful drug targets.

It has been suggested that using genetic association information is one of the best ways to identify such drug targets [2]. In recent years, a large number of highly powered GWAS studies have been published for numerous common traits (see for example [3] or [4]) and have yielded many candidate genes. Further potential targets can be identified by adding contextual data to the genetic associations, such as genes involved in similar biological processes [5, 6]. Biological networks and biological pathways can be used as a source of contextual data.

Biological networks are widely used in bioinformatics and can be constructed from multiple data sources, ranging from macromolecular interaction data collected from the literature [7] to correlation of expression in transcriptomics or proteomics samples of interest [8]. A large number of interaction network resources have been made available over the years, many of which are now in the public domain, combining thousands of interactions in a single location [9, 10]. They are based on three different fundamental types of data: (1) data-driven networks such as those built by WGCNA [8] for co-expression; (2) interactions extracted from the literature using a human curation process as exemplified by IntAct [11] or BioGRID [12]; and (3) interactions extracted from the literature using text mining approaches [13].

On the other hand, a plethora of network analysis algorithms are available for extracting useful information from such large biological networks in a variety of contexts. Algorithms range in complexity from simple first-neighbour approaches, where the direct neighbours of a gene of interest are assumed to be implicated in similar processes [14], to machine learning (ML) algorithms designed to learn from the features of the network to make more useful biological predictions (e.g. [15]).

One broad family of network analysis algorithms are the so-called Network Propagation approaches [16], used in contexts such as protein function prediction [17], disease gene identification [16] and cancer gene mutation identification [18]. In this paper, we perform a systematic review of the usefulness of network analysis methods for the purpose of identification of disease genes susceptible of being drug targets. Claims that such methods are helpful in that context have been made on numerous

¹occasions but a comprehensive validation study is lacking. One major challenge in ¹
²doing such a study is that it is not straightforward to define a list of known disease ²
³genes to be used for this purpose. ³

⁴ To address this, the Open Targets collaboration has been setup between phar- ⁴
⁵maceutical companies and public institutions to collect information on known drug ⁵
⁶targets and to help identify new ones [19]. A dedicated internet platform provides ⁶
⁷a free-to-use accessible resource summarising known data on gene-disease relation- ⁷
⁸ships from a number of data sources (e.g. known released drugs, genetic associations ⁸
⁹from GWAS, etc) [19]. ⁹

¹⁰ The purpose of this work is to quantify the performance of network-based meth- ¹⁰
¹¹ods to prioritise novel targets, using various networks and validation schemes, and ¹¹
¹²aiming at a faithful reflection of a realistic drug development scenario. We select ¹²
¹³a number of network approaches that are representative of several classes of algo- ¹³
¹⁴rithms, and test their ability to recover known disease genes by cross-validation. ¹⁴
¹⁵

¹⁶ We benchmark multiple definitions of disease genes, computational methods, bi- ¹⁶
¹⁷ological networks, validation schemes and performance metrics. We account for all ¹⁷
¹⁸possible combinations of such factors and derive guidelines for future disease gene ¹⁸
¹⁹identification studies. ¹⁹

²⁰

²¹**Results** ²¹

²²**Benchmark framework** ²²

²³ Our general approach, summarised in figure 1, consisted in using a biological net- ²³
²⁴work and a list of genes with prior disease-association scores as input to a network ²⁴
²⁵propagation approach. We used three cross-validation schemes -two take into ac- ²⁵
²⁶count protein complexes- in which some of the prior disease-association scores are ²⁶
²⁷hidden. The desired output was a new ranking of genes in terms of their association ²⁷
²⁸scores to the disease. Such ranking was compared to the known disease genes in ²⁸
²⁹the validation fold using several performance metrics. Given the amount of design ²⁹
³⁰factors and comparisons, the metrics were analysed through explorative additive ³⁰
³¹models (see Methods). Alternatively, we provide plots on the raw metrics in the ³¹
³²supplement, stratified by method in figures S10 and S11 or by disease in figures S12 ³²
³³and S13. ³³

We considered 2 metrics (AUROC and top 20 hits) and 2 input types (known drug¹
target genes and genetically associated genes), resulting in a total of 4 combinations,²
each described through an additive main effect model. Another 4 metrics were³
explored and can be found in the supplement (figure S17 and tables S6, S7).⁴

Interactions were explored, but they did not provide any added value for the⁵
extra complexity (see figure S18 from the supplement). The metrics used were the⁶
dependent variables, while the regressors included the prediction method, the CV⁷
scheme, the network and the disease.⁸

Performance using known drug targets as input¹⁰

Figure 2 describes the additive models for AUROC and top 20 hits, and using known¹¹
drug targets as input. Figure 3 contains their predictions for each method, network¹²
and cross validation scheme with 95% confidence intervals, averaged over diseases.¹³
The models are complex and we therefore review each main effect separately.¹⁴

AUPRC (quasi-binomial model) and top 20 hits (quasi-poisson) behave alike, as¹⁵
can be observed by their similar ranking of model estimates in Figure 2. For inter-¹⁶
pretability within real scenarios, only top 20 hits is shown in the main body. The¹⁷
standard AUROC (quasi-binomial) clearly led to different conclusions and is kept¹⁸
throughout the results section for comparison. The remaining metrics (AUPRC,¹⁹
pAUROC 5%, pAUROC 10% and top 100 hits) result in similar method prioritisa-²⁰
tions as top 20 hits, see figure S17. Detailed models can be found in the supplement,²¹
indexed by tables S6 and S7.²²

Comparing cross-validation schemes²³

Whether protein complexes were properly taken into account when performing the²⁴
cross-validation (see Methods) stood out as a key influence on the quality of predic-²⁵
tions: there was a dramatic reduction in performance for most methods when using²⁶
a complex-aware cross-validation strategy. For instance, method **rf** applied on the²⁷
STRING network dropped from almost 12 correct hits in the top 20 predicted dis-²⁸
ease genes when using our *classic* cross-validation scheme down to fewer than 4.5²⁹
when using either of our complex-aware cross-validation schemes. Likewise, table S5³⁰
from the supplement ratifies that only the *classic* cross validation splits complexes.³¹

Our data suggests that the performance drop when choosing the appropriate³²
validation strategy is comparable to the performance gap of competitive methods³³

¹versus a simple neighbour-voting baseline (see figure 2). This highlights the im-¹
²portance of carefully controlling for this bias when estimating the performance of²
³network-based disease gene prediction methods. Overall, the *classic* cross-validation³
⁴scheme gave biased estimates in our dataset, whereas our *block* and *representative*⁴
⁵cross-validation schemes had similar effects on the prediction performance. Method⁵
⁶ranking was independent of the cross validation choice thanks to the use of an addi-⁶
⁷tive model. And since both the *block* and *representative* schemes make theoretical⁷
⁸sense, we chose to focus on results from the block scheme in the rest of this study.⁸

⁹ *Comparing networks* ¹⁰

¹¹We found that using STRING as opposite to OmniPath improved overall perfor-¹¹
¹²mance of network-based disease gene prediction methods. Our models for top 20¹²
¹³hits quantified this effect as noticeable although less important than that of the¹³
¹⁴cross validation strategy. For reference, method **rf** obtains about 3 true hits under¹⁴
¹⁵both complex-aware strategies in OmniPath. It has been previously shown that the¹⁵
¹⁶positive effect on predictive power of having more interactions and coverage in a¹⁶
¹⁷network can outweigh the negative effect of increased number of false positive in-¹⁷
¹⁸teractions [20], which is in line with our findings. The authors also report STRING¹⁸
¹⁹among the best resources to discover disease genes, which is a finding we reproduce¹⁹
²⁰here. ²⁰

²¹We focus on the STRING results in the rest of the text. ²¹

²² *Comparing methods* ²²

²³Having identified the optimal cross-validation scheme and network for our bench-²³
²⁴mark in the previous sections, we quantitatively compared the performance of the²⁴
²⁵different methods. ²⁵

²⁶First, network topology alone had a slight predictive power, as method **pr** (PageR-²⁶
²⁷ank approach that ignores the input gene scores) showed better performance than²⁷
²⁸the **random** baseline under all the metrics. The randomised diffusion **randomraw**²⁸
²⁹lied between **random** and **pr** in performance. Both facts support the existence of an²⁹
³⁰inherent network topology-related bias among the positives that benefits diffusion-³⁰
³¹based methods. ³¹

³²Second, the basic GBA approach from EGAD had an advantage over the input-³²
³³naïve baselines **pr**, **randomraw** and **random**. It also outperformed prioritizing genes³³

¹using other Open Targets data stream scores such as genes associated to disease¹
²from pathways or from the literature (see supplement, table S19).²
³Most diffusion-based and ML-based methods outperformed EGAD. Results from³
⁴top 20 hits suggest using **rf** for the best performance followed by, in order: **raw** and⁴
⁵**bagsvm**, **z** and **svm** (main models panel in Figure 6).⁵
⁶To formally test the differences between methods, we carried a Tukey's multi-⁶
⁷ple comparison test on the model coefficients (Figure 5) as implemented in the R⁷
⁸package multcomp [21]. Although such differences were in most cases statistically⁸
⁹significant, even with such a strong multiplicity adjustment, their actual effect size⁹
¹⁰or magnitude can be modest in practice, see Figures 3 and 6.¹⁰
¹¹The ranking of methods was similar when using the metrics AUPRC, pAUROC¹¹
¹²and top *k* hits (see supplement, figure S17) and is only intended to be a general¹²
¹³reference, given the impact of the problem definition, cross validation scheme and¹³
¹⁴the network choice.¹⁴
¹⁵With AUROC on the other hand, **rf** lost its edge whilst most diffusion-based¹⁵
¹⁶and ML-based methods appeared technically tied. Despite its theoretical basis, in-¹⁶
¹⁷terpretability and widespread use in similar benchmarks, these results support the¹⁷
¹⁸assertion that AUROC is a sub-optimal choice in drug discovery practical scenarios.¹⁸
¹⁹Figure 4 further shows how the different methods compare with one another. Dis-¹⁹
²⁰tances between each pair of method in terms of their top 100 novel predictions were²⁰
²¹represented graphically. From this we observe that the supervised bagged SVM ap-²¹
²²proach (**bagsvm**) behaves similarly to the simple diffusion approach (**raw**), reflecting²²
²³the fact that they use the same kernel. We also observe that diffusion approaches do²³
²⁴not necessarily produce similar results (compare for example **raw** and **z**). And that²⁴
²⁵interestingly, methods EGAD (arguably one of the simplest) and COSNet (arguably²⁵
²⁶one of the most complex) seemed to result in similar predictions. Fully supervised²⁶
²⁷and semi-supervised approaches largely group in the top right hand quadrant of the²⁷
²⁸STRING plot away from diffusion methods, possibly showing some shared greater²⁸
²⁹potential for “learning effect” with the larger network.²⁹
³⁰Interestingly, when comparing overall performances shown in figure 6 with the³⁰
³¹prediction differences from the MDS plot (figure 4), it appears that the better³¹
³²performing methods may be doing well for different reasons as they do not occur³²
³³within the same region of the plot (e.g. **rf** and **raw**). MDS plots on the eight possible³³

¹combinations of network, input type and inclusion of seed genes are displayed in ¹
²the supplementary figures [S15](#) and [S16](#). ²

³ Regarding the STRING network and the block validation scheme, we fitted six ³
⁴additive models (one per metric) to the known drug target data (see supplement, ⁴
⁵table [S7](#)) and prioritised the methods (reduced models in figure [6](#)). These reduced ⁵
⁶models better described this particular scenario, as they were not forced to fit the ⁶
⁷trends in all networks and validation schemes in an additive way. Considering the ⁷
⁸top 20 hits, **rf** and **svm** were the optimal choices, followed by **wsld** and **knn**. ⁸

⁹ ⁹
¹⁰*Comparing diseases* ¹⁰

¹¹We next examine performance by disease. The top 20 hits model in figure [2](#) shows ¹¹
¹²that allergy (the figure's baseline reference), ulcerative colitis and rheumatoid arthri- ¹²
¹³tis (group I) are the diseases for which prediction of disease genes was worst, whereas ¹³
¹⁴cardiac arrhythmia, Parkinson's disease, stroke and multiple sclerosis (group II) are ¹⁴
¹⁵those for which it was best. As shown in figure [7](#), group I diseases had fewer known ¹⁵
¹⁶disease genes and lower modularity compared to group II diseases. ¹⁶

¹⁷ Prediction methods worked better when more known disease genes were available ¹⁷
¹⁸as input in the network, with two possible underlying reasons being the greater ¹⁸
¹⁹data availability to train the methods, and the natural bias of top 20 hits towards ¹⁹
²⁰datasets with more positives. Likewise, a stronger modularity within disease genes ²⁰
²¹justifies the guilt-by-association principle and led to better performances. In turn, ²¹
²²the number of genes and the modularity were positively correlated, see supplement, ²²
²³figure [S14](#). ²³

²⁴*Performance using genetic associations as input* ²⁴

²⁵ Using genetically associated genes as input to a prediction approach to find known ²⁵
²⁶drug targets mimicked a realistic scenario where novel genetic associations are ²⁶
²⁷screened as potential targets. However, inferring known drug targets through the ²⁷
²⁸indirect genetic evidence posed problems to prediction strategies, especially those ²⁸
²⁹based on machine learning. Learning is done using one class of genes in order to pre- ²⁹
³⁰dict genes that belong to another class, and the learning space suffers from intrinsic ³⁰
³¹uncertainties in the genetic associations to disease. ³¹

³² Consequently, we observed a major performance drop on all the prioritisation ³²
³³methods: using any network and cross-validation scheme, the predicted top 20 hits ³³

¹were practically bounded by 1. This was more pronounced on supervised machine¹
²learning-focused strategies, as **rf** and **svm** lost their edge on diffusion-based strate-²
³gies. The fact that the genetic associations of the validation fold were hidden further³
⁴hindered the predictions and can be a cause of our pessimistic performance esti-⁴
⁵mates. 5

⁶
⁷*Comparing cross-validation schemes* 7

⁸For reference, we also ran all three cross validation schemes on the genetic data to⁸
⁹quantify and account for complex-related bias. The models confirm that, contrary⁹
¹⁰to the drugs-related input, the differences between the results for the different cross-¹⁰
¹¹validation schemes were rather modest. For example, method **raw** with the STRING¹¹
¹²network attains 0.59-0.64, 0.50-0.54 and 0.37-0.40 hits in the top 20 under the¹²
¹³classical, block and representative cross-validation strategies. The slightly larger¹³
¹⁴negative effect on top 20 hits observed with the representative scheme is expected¹⁴
¹⁵because the number of positives that act as validation decreased and this metric¹⁵
¹⁶is biased by the class imbalance. The agreement between method ranking using¹⁶
¹⁷AUPRC and top 20 hits was less consistent, possibly due to the performance drop,¹⁷
¹⁸whilst AUROC again yielded quite a different ranking. Further data can be found¹⁸
¹⁹in the supplement, tables [S15](#) and [S16](#). 19

²⁰
²¹*Comparing networks* 21

²²The change in performance for using the OmniPath network instead of the filtered²²
²³STRING network was also limited. For AUROC the effect was negative, whereas for²³
²⁴the top 20 hits metric the performance improved. Method **raw** changed from 0.50-²⁴
²⁵0.54 top 20 hits in STRING to 0.61-0.66 in OmniPath under the block validation²⁵
²⁶strategy. 26

²⁷*Comparing methods* 27

²⁸To be consistent with the drugs section, we take as reference the block cross-²⁸
²⁹validation strategy and the STRING network. 29

³⁰The baseline approach **pr** that effectively makes use of the network topology alone³⁰
³¹proved difficult to improve upon, with 0.43-0.47 expected true hits in the top 20.³¹
³²Methods **raw** and **rf** respectively achieved 0.50-0.54 and 0.23-0.26 – although sig-³²
³³nificant, the difference in practice would be minimal. The best performing method³³

¹was mc with 0.65-0.7 hits. All the performance estimates can be found in the sup-¹
²plement, table [S16](#). To give an idea of the effort that would be required in a realistic²
³setting to find novel disease genes, the number of correct hits in the top 100 hits³
⁴was 3.29-3.45 with the best performing method (in this case, **ppr**), against 2.25-2.38⁴
⁵of **pr**.⁵

⁶ Two main conclusions can be drawn from these results. First, the network topology⁶
⁷baseline retained some predictive power upon which most diffusion-based methods,⁷
⁸as well as machine-learning approaches **COSNet** and **bagsvm**, only managed to add⁸
⁹minor improvements, if any. Second, drug targets could still be found by combining⁹
¹⁰network analysis and genes with genetic associations to disease, but with a substan-¹⁰
¹¹tially lower performance and with a marginal gain compared to a baseline approach¹¹
¹²that would only use the network topology to find targets (e.g. by screening the most¹²
¹³connected genes in the network).¹³

¹⁴ It is worth noting that gene-disease genetic association scores themselves have¹⁴
¹⁵drawbacks and that better prediction accuracy could result as genetic association¹⁵
¹⁶data improves.¹⁶

¹⁸**Conclusions**¹⁸

¹⁹We performed an extensive analysis of the ability of network-based approaches to¹⁹
²⁰identify novel disease genes. We exhaustively explored the effect of different factors²⁰
²¹including the biological network, the definition of disease genes, and the statistical²¹
²²framework being used to evaluate methods performance. We show that carefully²²
²³choosing an appropriate cross-validation framework and suitable performance met-²³
²⁴ric have an important effect in evaluating the utility of these methods.²⁴

²⁵ Our main conclusion is that network-based drug target discovery seems effective,²⁵
²⁶reflecting the fact that drug targets tend to cluster within the network. This in²⁶
²⁷turn may of course be due to the fact that the scientific community has so far²⁷
²⁸been focusing on testing the same proven mechanisms. In a strict cross-validation²⁸
²⁹setting, we found that even the most basic guilt-by-association method was useful,²⁹
³⁰with ~ 2 correct hits in its top 20 predictions, compared to ~ 0.1 when using a³⁰
³¹random ranking. The best diffusion based algorithm improved that figure to ~ 3.75 ,³¹
³²and the best overall performing method was a random forest classifier on network-³²
³³based features (~ 4.4 hits). Leading approaches can be notably different in terms³³

¹of their top predictions, suggesting potential complementarity. We found a better¹
²performance when using a network with more coverage at the expense of more false²
³positive interactions. In a more conservative network, random forest performance³
⁴dropped to ~3.1 hits. Comparing performance on different diseases shows that the⁴
⁵more known target genes, and the more clustered these are in the network, the⁵
⁶better the performance of network-based approaches for finding novel targets for it.⁶
⁷We also explored the prediction of known drug target genes by seeding the network⁷
⁸with an indirect data stream, in particular, genetic association data. Here, the best⁸
⁹performing methods were diffusion-based and presented a statistically significant,⁹
¹⁰but marginal, improvement over approaches that only look at network connectivity.¹⁰
¹¹We conclude that network propagation methods can help identify novel disease¹¹
¹²genes, but that the choice of the input network and the seed scores on the genes¹²
¹³needs careful consideration. Based on our approach and endorsed benchmarks, we¹³
¹⁴recommend the use of methods employing representations of diffusion-based in-¹⁴
¹⁵formation (the MashUp network-based features and the diffusion kernels), namely¹⁵
¹⁶random forest, the support vector machine variants, and raw diffusion algorithms¹⁶
¹⁷for optimal results.¹⁷

¹⁹Methods¹⁹

²⁰Selection of methods for investigation²⁰

²¹Algorithms were selected for validation based on the following criteria:²¹

- ²²1 published in a peer-reviewed journal, with evidence of improved performance²²
²³in gene disease prediction relative to contenders,²³
- ²⁴2 implemented as a well documented open source package, that is efficient, ro-²⁴
²⁵bust and usable within a batch testing framework,²⁵
- ²⁶3 directly applicable for gene disease identification from a single gene/protein²⁶
²⁷interaction network, without requiring fundamental changes to the approach²⁷
²⁸or additional annotation information and²⁸
- ²⁹4 capable of outputting a ranked list of individual genes (as opposed to gene²⁹
³⁰modules for example).³⁰

³¹In addition, we selected methods that were representative of a diverse panel of³¹
³²algorithms, including diffusion-based approaches, supervised learning approaches,³²
³³and a number of baseline approaches (see table 1).³³

¹Testing framework, algorithms and parameterisation 1

²All tests and batch runs were set-up and conducted using the R statistical pro-²

³gramming language [22]. When no R package was available, the methodology was³

⁴re-implemented, building upon existing R packages whenever possible. Standard⁴

⁵R machine learning libraries were used to train the support vector machine and⁵

⁶random forest classifiers. Only the MashUp algorithm [23] required feature gener-⁶

⁷ation outside of the R environment, using the Matlab code from their publication.⁷

⁸Further details on the methods implementation can be found in the supplement,⁸

⁹section “Method details”. 9

¹⁰EGAD [24], a pure neighbour-voting approach, was used here as a baseline com-¹⁰

¹¹parator. 11

¹²Diffusion (propagation) methods are central in this study. We used the random¹²

¹³walk-based personalised PageRank [25], previously used in similar tasks [26], as¹³

¹⁴implemented in *igraph* [27]. The remaining diffusion-based methods were run on¹⁴

¹⁵top of the regularised Laplacian kernel [28], computed through *diffuStats* [29]. We¹⁵

¹⁶included the classical diffusion **raw**, a weighted approach version **gm** and two statis-¹⁶

¹⁷tically normalised scores (**mc** and **z**), as implemented in *diffuStats*. In the scope of¹⁷

¹⁸positive-unlabelled learning [30, 31], we included the kernelised scores **knn** and the¹⁸

¹⁹linear decayed **wsld** from RANKS [32]. Closing this category, we implemented the¹⁹

²⁰bagging Support Vector Machine approach from ProDiGe1 [33], here **bagsvm**. 20

²¹Purer ML-based methods were also included. On one hand, network-based features²¹

²²were generated using MashUp [23] and two classical classifiers were fitted to them,²²

²³based on *caret* [34] and *mlr* [35]. These are **svm**, the Support Vector Machine as²³

²⁴implemented in *kernlab* [36], and **rf**, the Random Forest found in the *randomForest*²⁴

²⁵package [37]. On the other hand, we tried the parametric Hopfield recurrent neural²⁵

²⁶network classifier in the COSNet R package [38, 39]. 26

²⁷Finally, we defined three naive baseline methods: (1) **pr**, a classic problem-naïve²⁷

²⁸‘non-personalised’ PageRank implementation where input scores on the genes are²⁸

²⁹ignored; (2) **randomraw**, which applies the **raw** diffusion approach to randomly per-²⁹

³⁰mutated input scores on the genes; and (3) **random**, a uniform re-ranking of input³⁰

³¹genes without any network propagation. The inclusion of **pr** and **randomraw** al-³¹

³²lowed us to quantify the predictive power of the network topology alone, without³²

³³any consideration for the input scores on the genes. 33

¹Biological networks

²The biological network used in the validation is of critical importance as current²
³network resources contain both false positive and false negative interactions, and³
⁴these will affect any subsequent predictions [20].⁴

⁵Here, we used two human networks with different general properties, one more⁵
⁶likely to contain false positive interactions (STRING [40]), and another more con-⁶
⁷servative (OmniPath [41]), to test the effect of the network itself on network prop-⁷
⁸agation performance. We further filtered STRING [40] to retain only a subset of⁸
⁹interactions. Having tested several filters, we settled upon high-confidence inter-⁹
¹⁰actions (combined score > 700) with some evidence from the “Experiments” or¹⁰
¹¹“Databases” data sources (see supplement, table S2). Applying these filters and¹¹
¹²taking the largest connected component resulted in a connected network of 11,748¹²
¹³nodes and 236,963 edges. Edges were assigned weights between 0 and 1 by rescaling¹³
¹⁴the STRING combined score.¹⁴

¹⁵We did not filter the OmniPath network [41]. After removing duplicated edges¹⁵
¹⁶and taking the largest connected component, the OmniPath network contained¹⁶
¹⁷8,580 nodes and 42,145 unweighted edges.¹⁷

¹⁹Disease gene data

²⁰We used the Open Targets platform [19] to select known disease-related genes. In²⁰
²¹this analysis we define disease-related genes are those reported in Open Targets as²¹
²²being the target of a known drug against the disease of interest, or as those with²²
²³a genetic association of sufficient confidence with the disease. Associations were bi-²³
²⁴narised: any non-zero drugs-related association was considered positive, implying²⁴
²⁵that the methods would predict genes on which a drug has been essayed, regardless²⁵
²⁶of whether the drug was eventually approved. Likewise, only genetic associations²⁶
²⁷with an Open Targets score above 0.16 (see supplement, figure S1) were consid-²⁷
²⁸ered positive. We considered exclusively common diseases with at least 1,000 Open²⁸
²⁹Targets associations, of which a minimum of 50 could be based on known released²⁹
³⁰drugs and 50 on genetic associations, in order to avoid empty folds in the nested³⁰
³¹cross-validations. By applying these filters, we generated a list of phenotypes and³¹
³²diseases which we then manually curated to remove cancers (where causal genetic³²
³³mechanisms can differ from those of other common diseases), non-disease pheno-³³

¹type terms (e.g. “body weight and measures”) as well as vague or broad terms¹
²(e.g. “cerebrovascular disorder” or “head disease”) and infectious diseases. This left²
³22 diseases considered in this study (table 2). Further descriptive material on the³
⁴role of disease genes within the STRING network can be found in the section “De-⁴
⁵scriptive disease statistics in the STRING network” from the supplement. ⁵

⁶Validation strategies ⁶

⁷*Input Gene Scores* ⁷

⁸We used the binarised drug association scores and genetic association scores from ⁸
⁹Open Targets as input gene-level scores to seed the network propagation analyses ⁹
¹⁰(figure 8) and test their ability to recover known drug targets. With the first ap- ¹⁰
¹¹proach (subfigure (1) in figure 8), we tested the predictive power of current network ¹¹
¹²propagation methods for drug target identification using a direct source of evidence ¹²
¹³(known drug targets). In the second approach (subfigure (2) in figure 8), we assessed ¹³
¹⁴the ability of a reasonable but indirect source of evidence – genetic associations to ¹⁴
¹⁵disease – in combination with network propagation to recover known drug targets. ¹⁵
¹⁶
¹⁷

¹⁸*Metrics* ¹⁸

¹⁹Methods were systematically compared using standard performance metrics. The ¹⁹
²⁰Area under the Receiver Operating Characteristic curve (AUROC) is extensively ²⁰
²¹used in the literature for binary classification of disease genes [42], but can be ²¹
²²misleading in this context given the extent of the class imbalance between target and ²²
²³non-target genes [43]. We however included it in our benchmark for comparison with ²³
²⁴previous literature. More suitable measures of success in this case are Area under ²⁴
²⁵the Precision-Recall curve (AUPRC) [43] and partial AUROC (pAUROC) [44]. ²⁵
²⁶AUROC, AUPRC and pAUROC were computed with the precrec R package [45]. ²⁶
²⁷We also included top 20 hits, defined as the number of true positives in the top ²⁷
²⁸20 predicted genes (proportional to precision at 20). It is straightforward, intuitive ²⁸
²⁹and most likely to be useful in practice, such as a screening experiment where only ²⁹
³⁰a small number of predicted hits can be assayed. ³⁰

³¹We considered another 3 metrics, reported only in Supplement, i.e. partial AU- ³¹
³²ROC up to 5% FPR, partial AUROC up to 10% FPR, and number of hits within ³²
³³the top 100 genes. ³³

¹*Cross validation schemes and protein complexes* 1

²Standard (stratified) and modified k-fold cross-validation were used to estimate²
³the performance of network-based methods. Folds were based upon known drugs-³
⁴related genes, regardless of which type of input was used (see figure 8). Genes in⁴
⁵the training fold were negatively or positively labelled according to the input type,⁵
⁶whereas genes in the validation fold were left unlabelled. 6

⁷A fundamental challenge we faced when applying cross-validation to this prob-⁷
⁸lem was that known drug targets often consist in protein complexes, e.g. multi-⁸
⁹protein receptors. Drug-target associations typically have complex-level resolution.⁹
¹⁰The drug target data from Open Targets comes from ChEmbl [46], in which all the¹⁰
¹¹proteins in the targeted complex are labelled as targets. 11

¹²If left uncorrected, this could bias our cross-validation results: networks densely¹²
¹³connect proteins within a complex, random folds would frequently split positively¹³
¹⁴labelled complexes between train and validation, and therefore network-based meth-¹⁴
¹⁵ods would have an unfair advantage at spotting positive hits in the training folds.¹⁵
¹⁶In view of this, we benchmarked the methods under three cross validation strate-¹⁶
¹⁷gies: a standard cross validation (A) in line with usual practice and two (B, C)¹⁷
¹⁸complex-aware schemes (figure 9) addressing non-independence between folds when¹⁸
¹⁹the known drug targets act as input. 19

²⁰Strategy (A), called **classic**, was a regular stratified *k*-fold repeated cross-²⁰
²¹validation. We used *k* = 3 folds, averaging metrics over each set of folds, repeated²¹
²²25 times (see also figure 1). 22

²³Strategy (B), named **block**, performed a repeated cross validation while explicitly²³
²⁴preventing any complexes that contain disease genes to be split across folds. The key²⁴
²⁵point is that, where involved, shuffling was performed at the complex level instead²⁵
²⁶of the gene level – overlapping complexes that shared at least one known drug target²⁶
²⁷were merged into a larger pseudo-complex before shuffling. Fold boundaries were²⁷
²⁸chosen so that no complex was divided into two folds, while keeping them as close²⁸
²⁹as possible to those that would give a balanced partition, see figure 9. Nevertheless,²⁹
³⁰a limitation of this scheme is that it can fail to balance fold sizes in the presence³⁰
³¹of large complexes (see supplement, figure S9). For example, chronic obstructive³¹
³²pulmonary disease exhibited imbalanced folds, as 50 of the proteins involved belong³²
³³to the Mitochondrial Complex I 33

Strategy (C), referred to as **representative**, selected only a single representative¹
or prototype gene for each complex to ensure that gene information in a complex²
was not mixed between training and validation folds. In each repetition of cross val-³
idation, after merging the overlapping complexes, a single gene from each complex⁴
was chosen uniformly at random and kept as positive. The remaining genes from the⁵
complexes involved in the disease were set aside from the training and validation⁶
sets, in order (1) not to mislead methods into assuming their labels were negative⁷
in the training phase, and (2) not to overestimate (if set as positives) or penalise⁸
(if set as negatives) methods that ranked them highly, as they were expected to do⁹
so. This strategy kept the folds balanced, but at the expense of a possible loss of¹⁰
information by summarising each complex by a single gene at a time, reducing the¹¹
number of positives for training and validation.¹²

Additive performance models

For a systematic comparison between diseases, methods, cross-validation schemes¹⁶
and input types, we fitted an additive regression model to the performance metrics of¹⁷
each (averaged) fold from the cross-validation. The use of main effect models eased¹⁸
the evaluation of each individual factor while correcting for the other covariates.¹⁹
We modelled each metric f separately for each input type, not to mix problems of²⁰
different nature:²¹

$$f \sim \text{cv_scheme} + \text{network} + \text{method} + \text{disease}$$

We fitted dispersion-adjusted logistic-like *quasibinomial* distributions for the met-²⁵
rics AUROC, pAUROC and AUPRC and *quasipoisson* for top k hits. *The effect of*²⁶
*changing any of the four main effects is discussed in separate sub-sections in Re-*²⁷
sults, following the order from the formula above. After a data driven choice of²⁸
cross-validation scheme and network we fitted reduced models within them for a²⁹
more accurate description:³⁰

$$f \sim \text{method} + \text{disease}$$

¹*Qualitative methods comparison*

²The rankings produced by the different algorithms were qualitatively compared us-
³ing Spearman's footrule [47]. Distances were computed between all method ranking³
⁴pairs for each individual combination of disease, input type, network and for the⁴
⁵top N predicted genes, excluding the original seed genes. This part does not involve⁵
⁶cross validation – all known disease-associated genes were used for gene priori-⁶
⁷sations. Pairs of rankings could include genes uniquely ranked highly by a single⁷
⁸algorithm from the comparison, so mismatch counts (i.e. percentage mismatches)⁸
⁹between these rankings were also taken into account. Mismatches occur when a⁹
¹⁰gene features in the top N predictions of one algorithm and is missing from the¹⁰
¹¹corresponding ranking by another algorithm. A compact visualisation of distance¹¹
¹²matrices was obtained using a multi-view extension of MDS [48, 49, 50]. For this¹²
¹³we used the R package *multiview* [51] that generates a single, low-dimensional pro-¹³
¹⁴jection of combined inputs (disease, input and network).¹⁴

¹⁶**Competing interests**

¹⁷SB, DW, AG and BD are paid employees and shareholders of GlaxoSmithKline PLC. The commercial affiliation of
¹⁸SB, DW, AG and BD does not alter our adherence to BioMed Central policies.¹⁸

¹⁸**Availability of data and materials**

¹⁹The datasets supporting the conclusions of this article are available in <https://github.com/b2slab/genedise>.¹⁹

²⁰**Author's contributions**

²¹SP, SB, DW, and BD analysed and interpreted the data. AP, AG and BD helped supervise the project. All authors
²²provided critical feedback and helped shape the research, analysis and manuscript. All authors approved the final
²³version of this manuscript for publication.²²

²³**Acknowledgements**

²⁴AG would like to acknowledge Philippe Sanseau, Matt Nelson and John Whittaker for critical feedback on the
²⁵applicability of this research to drug discovery.²⁴

²⁶This work was supported by the Spanish Ministry of Economy and Competitiveness (MINECO) [TEC2014-60337-R
²⁷and DPI2017-89827-R to AP].²⁵

²⁸AP and SP thank for funding the Spanish Networking Biomedical Research Centre in the subject area of
²⁹Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN), initiative of Instituto de Investigación Carlos III
³⁰(ISCIII). SP thanks the AGAUR FI-scholarship programme.²⁶

²⁸**Author details**

²⁹¹B2SLab, Departament d'Enginyeria de Sistemes, Automàtica i Informàtica Industrial, Universitat Politècnica de
³⁰Catalunya, CIBER-BBN, 08028, Barcelona, Spain. ²Institut de Recerca Pediàtrica Hospital Sant Joan de Déu,
³¹08950, Esplugues de Llobregat, Barcelona, Spain. ³Computational Biology and Statistics, GSK, Gunnels Wood
³²Road, SG1 2NY Stevenage, United Kingdom. ⁴Research Statistics, GSK, Gunnels Wood Road, SG1 2NY
³³Stevenage, United Kingdom. ⁵GSK Vaccines, Rue de l'Institut 89, 1330 Rixensart, Belgium.²⁷

³²**References**

³³1. Scannell, J.W., Blanckley, A., Boldon, H., Warrington, B.: Diagnosing the decline in pharmaceutical r&d
efficiency. *Nature reviews Drug discovery* **11**(3), 191–200 (2012)³³

- 1 2. Nelson, M.R., Tipney, H., Painter, J.L., Shen, J., Nicoletti, P., Shen, Y., Floratos, A., Sham, P.C., Li, M.J., 1
- 2 Wang, J., Cardon, L.R., Whittaker, J.C., Sanseau, P.: The support of human genetic evidence for approved 2
- 3 drug indications. *Nature Genetics* **47**(8), 856–860 (2015). doi:[10.1038/ng.3314](https://doi.org/10.1038/ng.3314) 2
- 3 3. Verstockt, B., Smith, K.G., Lee, J.C.: Genome-wide association studies in crohn's disease: Past, present and 3
- 4 future. *Clinical & Translational Immunology* **7**(1), 1001 (2018). doi:[10.1002/cti2.1001](https://doi.org/10.1002/cti2.1001) 4
- 4 4. of the Psychiatric Genomics Consortium, S.W.G.: Biological insights from 108 schizophrenia-associated genetic 4
- 5 loci. *Nature* **511**(7510), 421–427 (2014). doi:[10.1038/nature13595](https://doi.org/10.1038/nature13595) 5
- 5 5. Jia, P., Zhao, Z.: Network-assisted analysis to prioritize GWAS results: principles, methods and perspectives. 5
- 6 *Human Genetics* **133**(2), 125–138 (2013). doi:[10.1007/s00439-013-1377-1](https://doi.org/10.1007/s00439-013-1377-1) 6
- 6 6. Boyle, E.A., Li, Y.I., Pritchard, J.K.: An expanded view of complex traits: From polygenic to omnigenic. *Cell* 6
- 7 **169**(7), 1177–1186 (2017). doi:[10.1016/j.cell.2017.05.038](https://doi.org/10.1016/j.cell.2017.05.038) 7
- 7 7. Orchard, S., Kerrien, S., Abbani, S., Aranda, B., Bhate, J., Bidwell, S., Bridge, A., Briganti, L., Brinkman, 7
- 8 F.S., Brinkman, F., Cesareni, G., Chatr-aryamontri, A., Chautard, E., Chen, C., Dumousseau, M., Goll, J., 8
- 9 Hancock, R.E., Hancock, R., Hannick, L.I., Jurisica, I., Khadake, J., Lynn, D.J., Mahadevan, U., Perfetto, L., 9
- 10 Raghunath, A., Ricard-Blum, S., Roechert, B., Salwinski, L., Stumpflen, V., Tyers, M., Uetz, P., Xenarios, I., 10
- 11 Hermjakob, H.: Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. 11
- 12 *Nat. Methods* **9**(4), 345–350 (2012) 11
- 12 8. Langfelder, P., Horvath, S.: WGCNA: an R package for weighted correlation network analysis. *BMC* 12
- 13 *Bioinformatics* **9**(1), 559 (2008). doi:[10.1186/1471-2105-9-559](https://doi.org/10.1186/1471-2105-9-559) 13
- 13 9. Razick, S., Magklaras, G., Donaldson, I.M.: iRefIndex: A consolidated protein interaction database with 13
- 14 provenance. *BMC Bioinformatics* **9**(1), 405 (2008). doi:[10.1186/1471-2105-9-405](https://doi.org/10.1186/1471-2105-9-405) 14
- 14 10. Szklarczyk, D., Morris, J.H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N.T., 14
- 15 Roth, A., Bork, P., Jensen, L.J., von Mering, C.: The STRING database in 2017: quality-controlled 15
- 16 protein–protein association networks, made broadly accessible. *Nucleic Acids Research* **45**(D1), 362–368 (2016). 16
- 17 doi:[10.1093/nar/gkw937](https://doi.org/10.1093/nar/gkw937) 16
- 17 11. Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., Duesbury, M., Dumousseau, M., 17
- 18 Feuermann, M., Hinz, U., *et al.*: The intact molecular interaction database in 2012. *Nucleic acids research* 18
- 19 **40**(D1), 841–846 (2011) 18
- 19 12. Chatr-Aryamontri, A., Oughtred, R., Boucher, L., Rust, J., Chang, C., Kolas, N.K., O'Donnell, L., Oster, S., 19
- 20 Theesfeld, C., Sellam, A., *et al.*: The biogrid interaction database: 2017 update. *Nucleic acids research* **45**(D1), 19
- 20 369–379 (2017) 20
- 21 13. Al-Aamri, A., Taha, K., Al-Hammadi, Y., Maalouf, M., Homouz, D.: Constructing genetic networks using 21
- 22 biomedical literature and rare event classification. *Scientific reports* **7**(1), 15784 (2017) 21
- 22 14. Piovesan, D., Giollo, M., Ferrari, C., Tosatto, S.C.E.: Protein function prediction using guilty by association 22
- 23 from interaction networks. *Amino Acids* **47**(12), 2583–2592 (2015). doi:[10.1007/s00726-015-2049-3](https://doi.org/10.1007/s00726-015-2049-3) 23
- 23 15. Re, M., Mesiti, M., Valentini, G.: A fast ranking algorithm for predicting gene functions in biomolecular 23
- 24 networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **9**(6), 1812–1818 (2012) 24
- 24 16. Cowen, L., Ideker, T., Raphael, B.J., Sharan, R.: Network propagation: a universal amplifier of genetic 24
- 25 associations. *Nature Reviews Genetics* **18**(9), 551–562 (2017). doi:[10.1038/nrg.2017.38](https://doi.org/10.1038/nrg.2017.38) 25
- 25 17. Sharan, R., Ulitsky, I., Shamir, R.: Network-based prediction of protein function. *Molecular Systems Biology* **3** 26
- 26 (2007). doi:[10.1038/msb4100129](https://doi.org/10.1038/msb4100129) 26
- 27 18. Leiserson, M.D.M., Vandin, F., Wu, H.-T., Dobson, J.R., Eldridge, J.V., Thomas, J.L., Papoutsaki, A., Kim, 27
- 28 Y., Niu, B., McLellan, M., Lawrence, M.S., Gonzalez-Perez, A., Tamborero, D., Cheng, Y., Ryslik, G.A., 28
- 29 Lopez-Bigas, N., Getz, G., Ding, L., Raphael, B.J.: Pan-cancer network analysis identifies combinations of rare 29
- 30 somatic mutations across pathways and protein complexes. *Nature Genetics* **47**(2), 106–114 (2014). 30
- 30 doi:[10.1038/ng.3168](https://doi.org/10.1038/ng.3168) 30
- 31 19. Koscielny, G., An, P., Carvalho-Silva, D., Cham, J.A., Fumis, L., Gasparyan, R., Hasan, S., Karamanis, N., 31
- 31 Maguire, M., Papa, E., *et al.*: Open targets: a platform for therapeutic target identification and validation. 31
- 32 *Nucleic acids research* **45**(D1), 985–994 (2016) 32
- 32 20. Huang, J.K., Carlin, D.E., Yu, M.K., Zhang, W., Kreisberg, J.F., Tamayo, P., Ideker, T.: Systematic evaluation 32
- 33 of molecular networks for discovery of disease genes. *Cell systems* **6**(4), 484–495 (2018) 33

- 1 21. Hothorn, T., Bretz, F., Hothorn, M.T.: The multcomp package (2009) 1
- 2 22. R Core Team: R: A language and environment for statistical computing (2016). R Foundation for Statistical 2
- 3 Computing 3
- 4 23. Cho, H., Berger, B., Peng, J.: Compact integration of multi-network topology for functional analysis of genes. 4
- 5 Cell systems 3(6), 540–548 (2016) 4
- 6 24. Ballouz, S., Weber, M., Pavlidis, P., Gillis, J.: Egad: ultra-fast functional analysis of gene networks. 5
- 7 Bioinformatics 33(4), 612–614 (2017). doi:[10.1093/bioinformatics/btX000](https://doi.org/10.1093/bioinformatics/btX000) 5
- 8 25. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. 6
- 9 Technical report, Stanford InfoLab (1999) 6
- 10 26. Jiang, B., Kloster, K., Gleich, D.F., Gribskov, M.: Aptrank: an adaptive pagerank model for protein function 7
- 11 prediction on bi-relational graphs. Bioinformatics 33(12), 1829–1836 (2017) 7
- 12 27. Csardi, G., Nepusz, T.: The igraph software package for complex network research. InterJournal **Complex 8**
- 13 **Systems**, 1695 (2006) 8
- 14 28. Smola, A.J., Kondor, R.: Kernels and regularization on graphs. In: Learning Theory and Kernel Machines, pp. 9
- 15 144–158. Springer, ??? (2003) 9
- 16 29. Picart-Armada, S., Thompson, W.K., Buil, A., Perera-Lluna, A.: An r package to compute diffusion-based 10
- 17 scores on biological networks: diffstats. Bioinformatics 32(18), 2872–2874 (2016). 10
- 18 doi:[10.1093/bioinformatics/btX000](https://doi.org/10.1093/bioinformatics/btX000) 11
- 19 30. Elkan, C., Noto, K.: Learning classifiers from only positive and unlabeled data, 213–220 (2008). ACM 12
- 20 31. Yang, P., Li, X.-L., Mei, J.-P., Kwok, C.-K., Ng, S.-K.: Positive-unlabeled learning for disease gene 13
- 21 identification. Bioinformatics 28(20), 2640–2647 (2012) 13
- 22 32. Valentini, G., Paccanaro, A., Caniza, H., Romero, A.E., Re, M.: Ranks: a flexible tool for node label ranking 14
- 23 and classification in biological networks. Artificial Intelligence in Medicine 61(2), 63–78 (2014) 14
- 24 33. Mordelet, F., Vert, J.-P.: ProDiGe: Prioritization Of Disease Genes with multitask machine learning from 15
- 25 positive and unlabeled examples. BMC bioinformatics 12(1), 389 (2011) 15
- 26 34. Kunn, M.: Building predictive models in r using the caret package. Journal of Statistical Software 28(5), 1–26 16
- 27 (2008) 16
- 28 35. Bischl, B., Lang, M., Kothhoff, L., Schiffner, J., Richter, J., Studerus, E., Casalicchio, G., Jones, Z.M.: mlr: 17
- 29 Machine learning in r. Journal of Machine Learning Research 17(170), 1–5 (2016) 17
- 30 36. Karatzoglou, A., Smola, A., Hornik, K., Zeileis, A.: kernlab – an S4 package for kernel methods in R. Journal of 18
- 31 Statistical Software 11(9), 1–20 (2004) 18
- 32 37. Liaw, A., Wiener, M.: Classification and regression by randomforest. R News 2(3), 18–22 (2002) 19
- 33 38. Frasca, M., Bertoni, A., Re, M., Valentini, G.: A neural network algorithm for semi-supervised node label 20
- 34 learning from unbalanced data. Bioinformatics 43(C), 84–98 (2013) 20
- 35 39. Bertoni, A., Frasca, M., Valentini, G.: Cosnet: a cost sensitive neural network for semi-supervised learning in 21
- 36 graphs. Lecture Notes in Computer Science 6911(4), 219–234 (2011). doi:[10.1007/978-3-642-21440-0_14](https://doi.org/10.1007/978-3-642-21440-0_14) 21
- 37 40. Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, 22
- 38 A., Santos, A., Tsafou, K.P., *et al.*: String v10: protein–protein interaction networks, integrated over the tree of 22
- 39 life. Nucleic acids research 43(D1), 447–452 (2014) 22
- 40 41. Türei, D., Korcsmáros, T., Saez-Rodriguez, J.: Omnipath: guidelines and gateway for literature-curated 23
- 41 signaling pathway resources. Nature methods 13(12), 966 (2016) 23
- 42 42. Lee, I., Blom, U.M., Wang, P.I., Shim, J.E., Marcotte, E.M.: Prioritizing candidate disease genes by 24
- 43 network-based boosting of genome-wide association data. Genome Research 21(7), 1109–1121 (2011). 24
- 44 doi:[10.1101/gr.118992.110](https://doi.org/10.1101/gr.118992.110) 24
- 45 43. Saito, T., Rehmsmeier, M.: The precision-recall plot is more informative than the ROC plot when evaluating 25
- 46 binary classifiers on imbalanced datasets. PLOS ONE 10(3), 0118432 (2015). 25
- 47 doi:[10.1371/journal.pone.0118432](https://doi.org/10.1371/journal.pone.0118432) 25
- 48 44. Jiang, Y., Metz, C.E., Nishikawa, R.M.: A receiver operating characteristic partial area index for highly sensitive 26
- 49 diagnostic tests. Radiology 201(3), 745–750 (1996) 26
- 50 45. Saito, T., Rehmsmeier, M.: Precrec: fast and accurate precision-recall and roc curve calculations in r. 27
- 51 Bioinformatics 33 (1), 145–147 (2017). doi:[10.1093/bioinformatics/btw570](https://doi.org/10.1093/bioinformatics/btw570) 27

1 46. Bento, A.P., Gaulton, A., Hersey, A., Bellis, L.J., Chambers, J., Davies, M., Krüger, F.A., Light, Y., Mak, L., 1
2 McGlinchey, S., *et al.*: The chEMBL bioactivity database: an update. *Nucleic acids research* **42**(D1), 1083–1090 2
(2014)
3 47. Spearman, C.: footrule for measuring correlation. *British Journal of Psychology* **2**(1), 89–108 (1906) 3
4 48. Mardia, K.V.: Some properties of classical multi-dimensional scaling. *Communications in Statistics-Theory and 4
Methods* **7**(13), 1233–1241 (1978)
5 49. Gower, J.C.: Some distance properties of latent root and vector methods used in multivariate analysis. 5
Biometrika **53**(3–4), 325–338 (1966)
6 50. Kanaan-Izquierdo, S., Ziyatdinov, A., Perera-Lluna, A.: Multiview and multifeature spectral clustering using 6
7 common eigenvectors. *Pattern Recognition Letters* **102**, 30–36 (2018) 7
8 51. Kanaan-Izquierdo, S.: multiview: A package with multiview clustering and dimensionality reduction methods. 8
<https://github.com/b2slab/multiview>
9 52. Team, B.C.: Bioconductor software project. *Bioconductor Open Source Software For Bioinformatics* 9
53. Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y.,
10 Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, 10
11 A.J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J.Y., Zhang, J.: Bioconductor: open software 11
development for computational biology and bioinformatics. *Genome Biology* **5**(R80) (2004)
12 54. Vandin, F., Upfal, E., Raphael, B.J.: Algorithms for detecting significantly mutated pathways in cancer. *Journal 12
of Computational Biology* **18**(3), 507–22 (2011)
13 55. Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C., Morris, Q.: Genemania: a real-time multiple association 13
network integration algorithm for predicting gene function. *Genome Biology* **9**(S4), 1–15 (2008) 14
14 56. Bersanelli, M., Mosca, E., Remondini, D., Castellani, G., Milanesi, L.: Network diffusion-based analysis of 14
15 high-throughput data for the detection of differentially enriched modules. *Scientific Reports* **6**(34841), 1–12 15
(2016)
16 57. Vandin, F., Upfal, E., Raphael, B.J.: Algorithms for detecting significantly mutated pathways in cancer. *IEEE 16
Signal Processing Magazine* **30**(4), 87–97 (2013) 17
18 58. Valentini, G., Armano, G., Frasca, M., Lin, J., Mesiti, M., Re, M.: Ranks: a flexible tool for node label ranking 18
and classification in biological networks. *Bioinformatics* **32**(18), 2872–2874 (2016)
19 59. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple 19
testing. *Journal of the royal statistical society. Series B (Methodological)*, 289–300 (1995)
20
21
22

21 **Figures**

23 **Tables**

25 **Table 1 List of methods included in this benchmark.** Method identifiers are shortened method names used throughout the text. Other columns are self-explanatory.

Method Identifier	Method Name	Method Class	Implementation	Reference
pr	PageRank with a uniform prior	Baseline	igraph (Bioconductor [52, 53] package)	[25]
random	Random	Baseline	R	(see text)
randomraw	Random Raw	Baseline	R	(see text)
EGAD	Extending Guilt by Association' by Degree	Baseline	EGAD (Bioconductor package)	[24]
ppr	Personalized PageRank	Diffusion	igraph (R package)	[26]
raw	Raw Diffusion	Diffusion	diffuStats (Bioconductor package)	[54]
gm	GeneMania-based weights	Diffusion	diffuStats (Bioconductor package)	[55]
mc	Monte Carlo normalised scores	Diffusion	diffuStats (Bioconductor package)	[56]
z	Z-scores	Diffusion	diffuStats (Bioconductor package)	[57]
knn	K nearest neighbours	Semi-supervised learning	RANKS (R package)	[58]
wsld	Weighted Sum with Linear Decay	Semi-supervised learning	RANKS (R package)	[58]
COSNet	COS Sensitive neural Network	Supervised learning	COSNet (R package)	[38]
bagsvm	Bagging SVM (based on ProDiGe1)	Supervised learning	kernlab (R package)	[33]
rf	Random Forest	Supervised learning	randomForest (R package) + Matlab (features)	[23]
svm	Support Vector Machine	Supervised learning	kernlab (R package) + Matlab (features)	[23]

Table 2 List of diseases included in this study. Diseases included in this study, with a minimum of 50 associated genes both in the known drug targets and the genetic categories (see text). The overlap between these two lists of genes showed a degree of dependence between these two Open Targets data streams for some of the diseases. P-values were calculated using Fisher's exact test and are reported without and with correction for false discovery rate (Benjamini and Hochberg [59]).

Disease	N(genetic)	N(drugs)	Overlap	P-value	FDR
allergy	112	57	1	4.22e-01	4.42e-01
Alzheimers disease	208	103	4	1.10e-01	1.42e-01
arthritis	174	188	6	6.08e-02	1.03e-01
asthma	105	80	6	7.77e-05	5.70e-04
bipolar disorder	117	148	3	1.83e-01	2.12e-01
cardiac arrhythmia	75	177	6	9.15e-04	3.36e-03
chronic obstructive pulmonary disease (COPD)	154	116	6	4.18e-03	1.31e-02
coronary heart disease	111	171	4	7.86e-02	1.24e-01
drug dependence	75	143	6	2.96e-04	1.30e-03
hypertension	66	188	2	2.85e-01	3.14e-01
multiple sclerosis	71	167	4	1.83e-02	4.03e-02
obesity	69	194	3	1.06e-01	1.42e-01
Parkinson's disease	55	145	0	1	1
psoriasis	131	105	7	1.68e-04	9.23e-04
rheumatoid arthritis	138	95	5	5.18e-03	1.42e-02
schizophrenia	410	163	17	5.44e-05	5.70e-04
stroke	90	156	3	1.18e-01	1.44e-01
systemic lupus erythematosus (lupus)	126	109	5	6.30e-03	1.54e-02
type I diabetes mellitus	87	106	3	4.39e-02	8.04e-02
type II diabetes mellitus	130	154	4	9.14e-02	1.34e-01
ulcerative colitis	136	51	7	1.81e-06	3.98e-05
unipolar depression	123	121	4	3.81e-02	7.63e-02

Additional Files

Additional file 1 — Supplement

This document contains complementary material that supports our claims in the main body. It includes topics such as descriptive statistics, topological properties of disease genes, raw metrics plots, method details, MDS plots, alternative performance metrics and further explicative models.

Additional file 2 — MDS_plots

Complementary single-disease MDS plots and distance matrices.

Additional file 3 — Interactions.html_viewer

Stand-alone viewer to explore models with interaction terms.

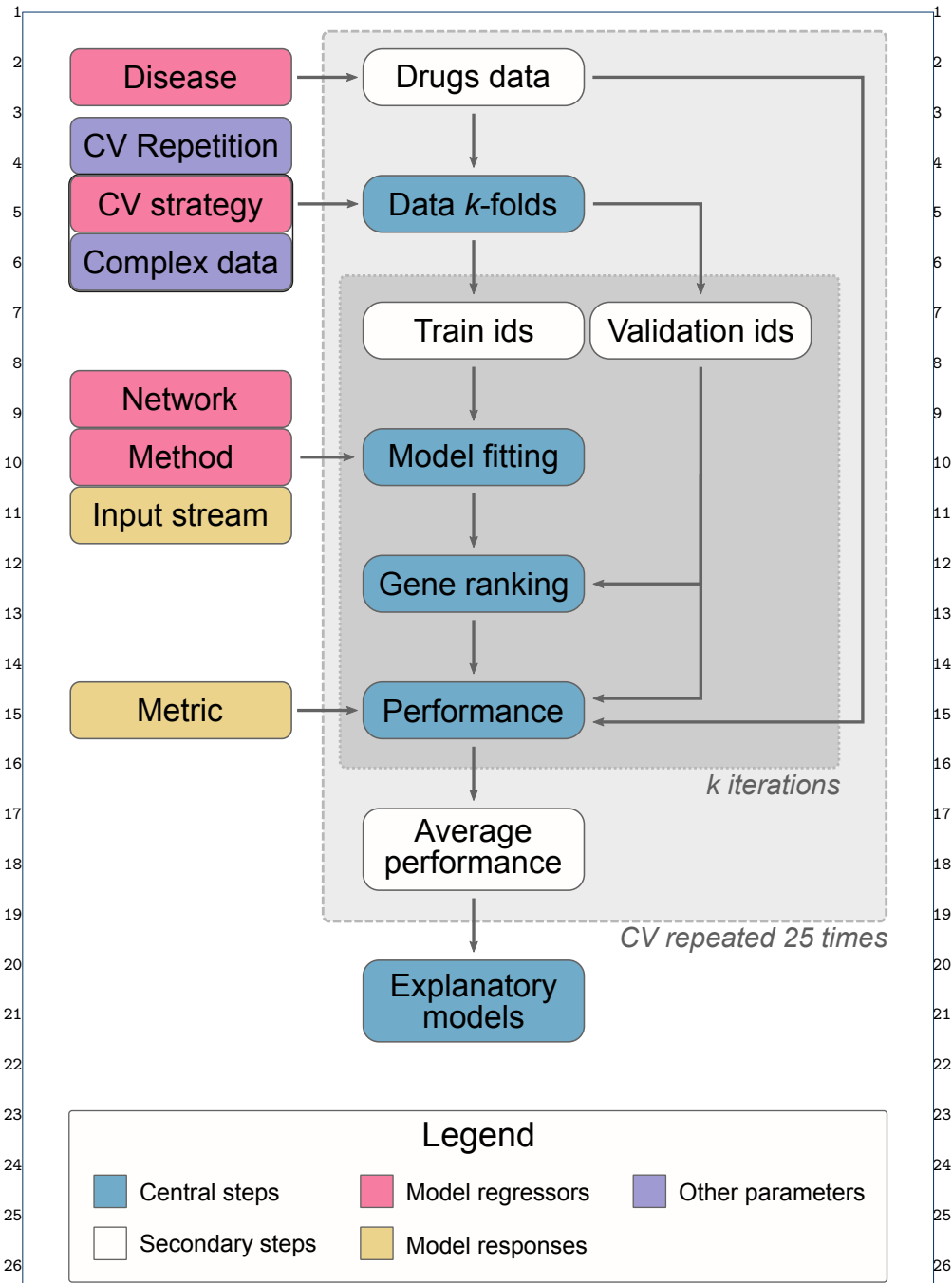


Figure 1 Benchmark overview. This work describes six performance metrics using two input streams (genetic association and drug-based genes) to predict drug-based genes for 22 common diseases. 3-fold cross validation (CV), repeated 25 times, was run under three CV strategies. The gene identifiers in each fold are determined using only the drugs data, regardless of the input. Two validation strategies are complex-aware and therefore needed this data to define the splits. 15 network-based methods (including 4 baselines) were evaluated, using two networks with different properties, by modelling their performance, averaged on every CV round. The explanatory models allowed hypothesis testing and a direct comparison between diseases, CV strategies, networks and methods.

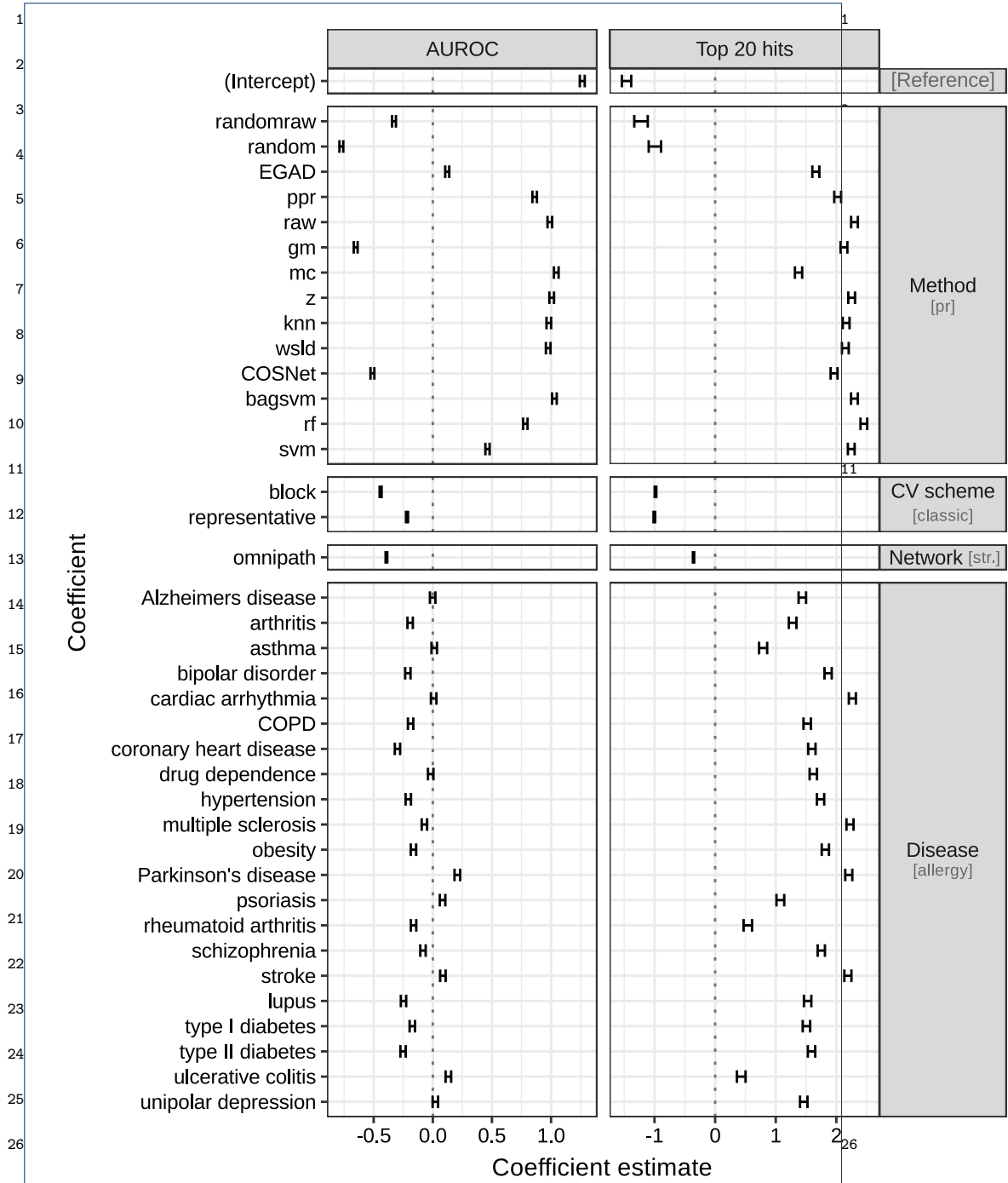


Figure 2 Additive models for AUROC and top 20 hits. Each column corresponds to a different model, whereas each row depicts the 95% confidence interval for each model coefficient. Rows are grouped by the categorical variable they belong to: method, cv scheme, network and disease. Each variable has a **reference level**, implicit in the intercept and specified in brackets: **pr** method, **classic** validation scheme, **STRING** network and **allergy**. Positive estimates improve performance over the reference levels, whereas negative ones reduce it. For example, the data suggest that method **rf** performs better than the baseline using both metrics, and is the preferred method using the top 20 hits. Switching from **STRING** to the **OmniPath** network, or from **classic** to **block** or **representative** cross-validation, has a negative effect on both performance metrics. Specific model estimates and confidence intervals can be found in the supplement, see tables [S8](#) and [S9](#).

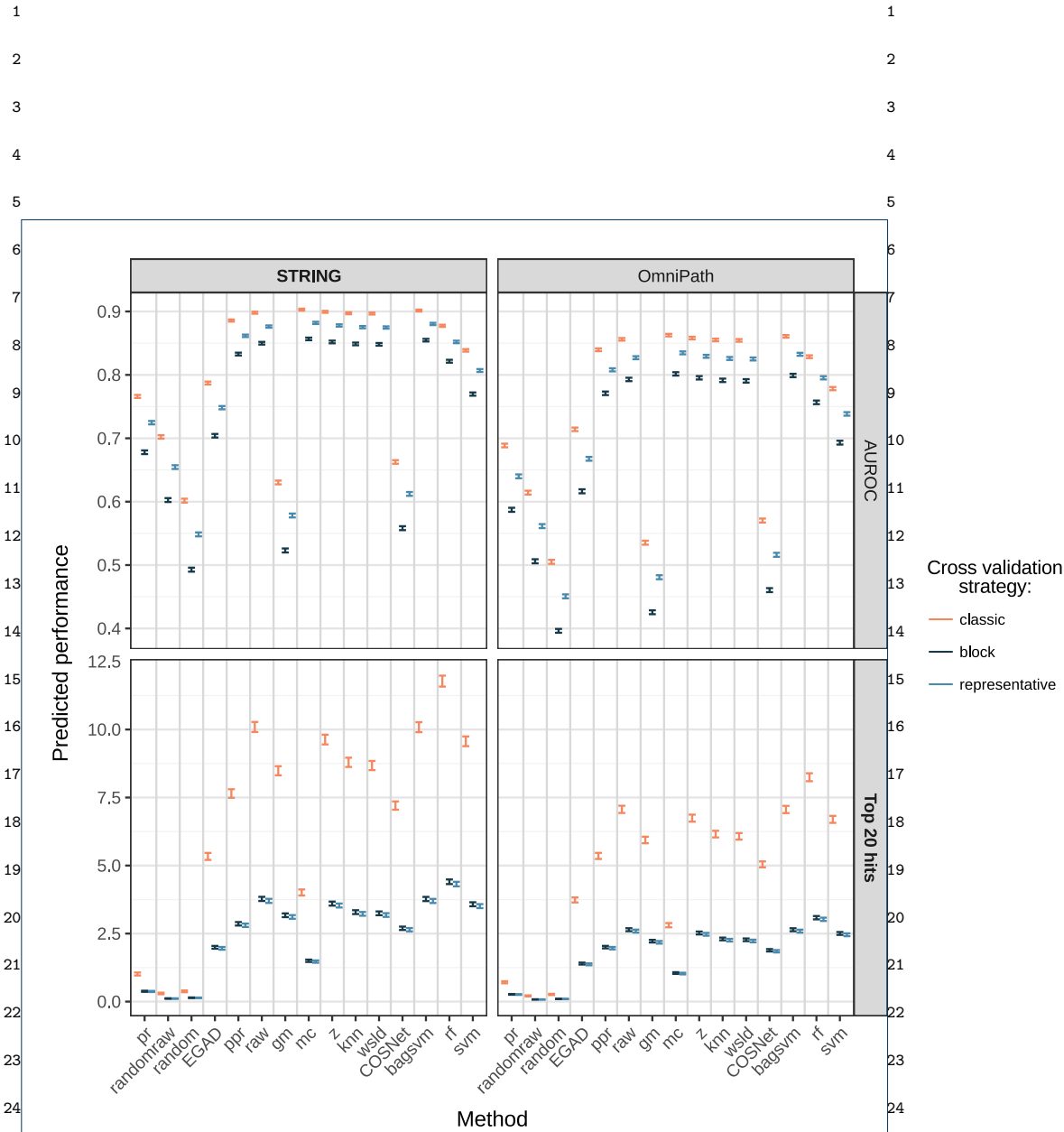
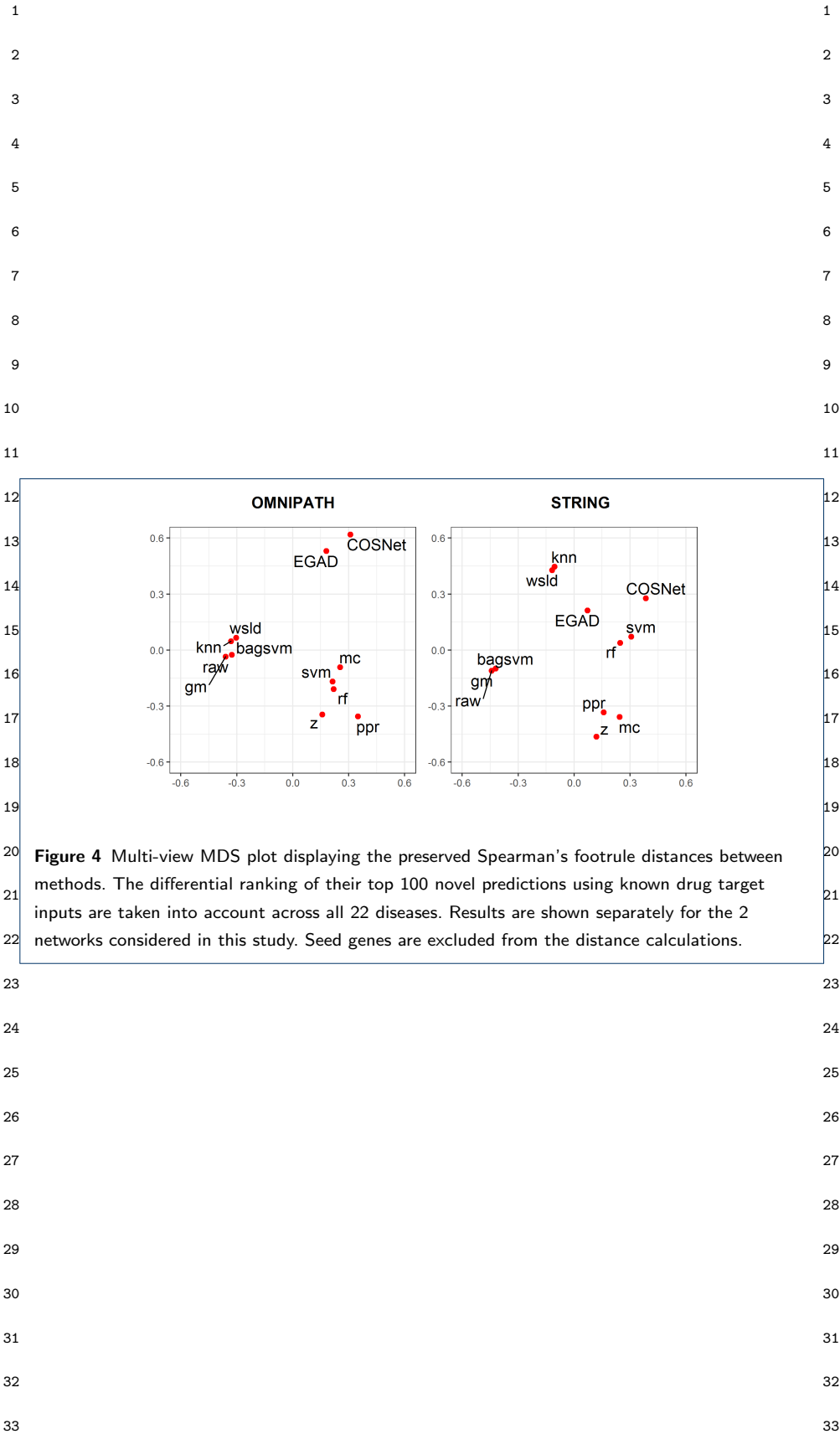


Figure 3 Performance predicted for AUROC and top 20 hits through the additive models. Each row corresponds to a different model and error bars depicts the 95% confidence interval of the additive model prediction, averaging over diseases. In bold, the main network (STRING) and metrics (AUPRC, top 20 hits). A table with the exact values can be found in the supplement, table S9.



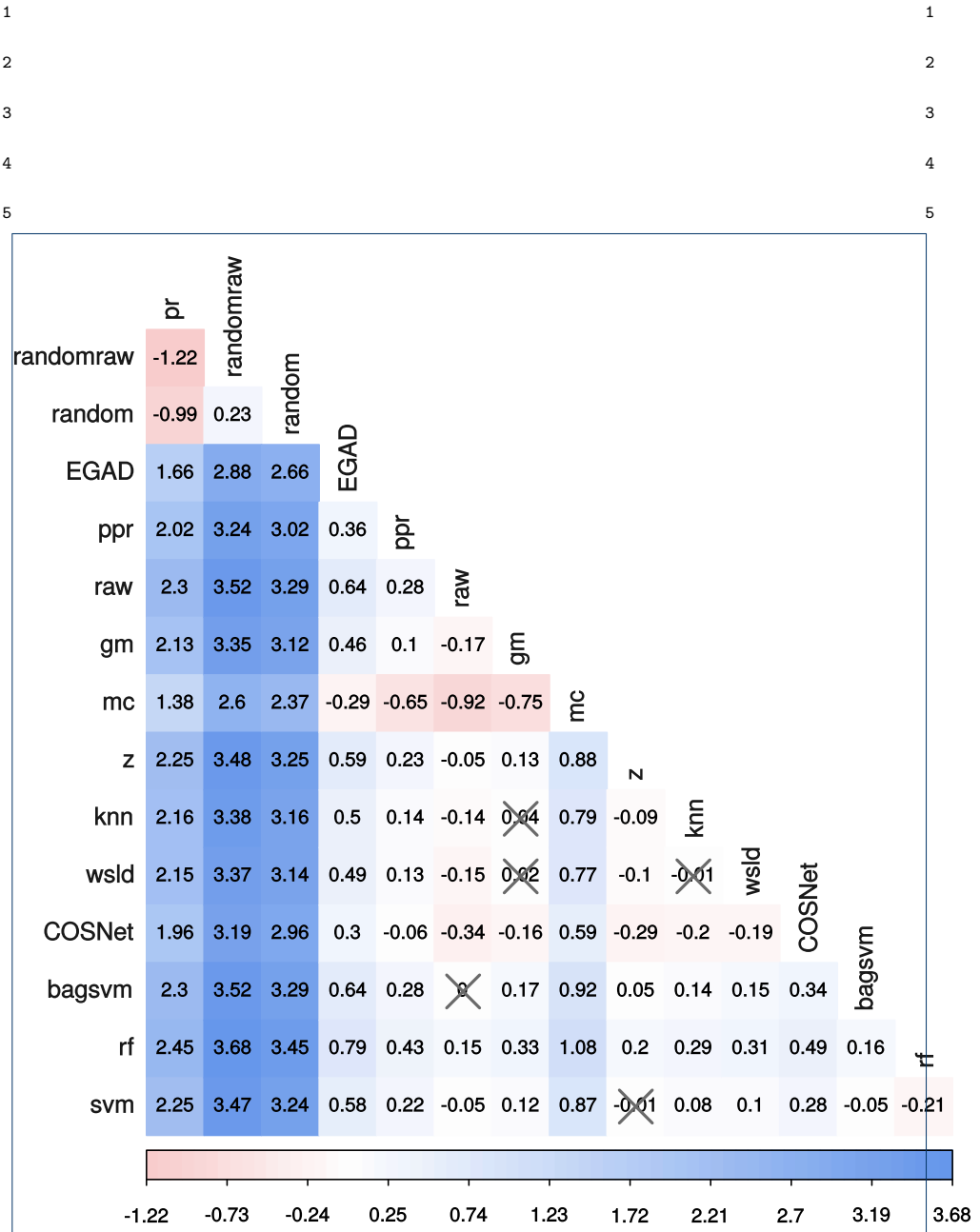


Figure 5 Pairwise contrasts on top 20 hits predicted by the quasipoisson model. Differences are expressed in the model space. Most of the pairwise differences are significant (Tukey's test, $p < 0.05$) – non-significant differences have been crossed out.

